Genome-Wide Association Study in Esophageal Cancer Using GeneChip Mapping 10K Array

Nan Hu,[1] Chaoyu Wang,[1] Ying Hu,[2] Howard H. Yang,[2] Carol Giffen,[5] Ze-Zhong Tang,[6] Xiao-Yu Han,[6] Alisa M. Goldstein,[4] Michael R. Emmert-Buck,[3] Kenneth H. Buetow,[2] Philip R. Taylor,[1] and Maxwell P. Lee[2]

[1]Cancer Prevention Studies Branch, [2]Laboratory of Population Genetics, and [3]Laboratory of Pathology, Center for Cancer Research; [4]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland; [5]Information Management Service, Inc., Silver Spring, Maryland; and [6]Shanxi Cancer Hospital, Taiyuan, Shanxi, People's Republic of China

Abstract

Whole genome association studies of complex human diseases represent a new paradigm in the postgenomic era. In this study, we report application of the Affymetrix, Inc. (Santa Clara, CA) high-density single nucleotide polymorphism (SNP) array containing 11,555 SNPs in a pilot case-control study of esophageal squamous cell carcinoma (ESCC) that included the analysis of germ line samples from 50 ESCC patients and 50 matched controls. The average genotyping call rate for the 100 samples analyzed was 96%. Using the generalized linear model (GLM) with adjustment for potential confounders and multiple comparisons, we identified 37 SNPs associated with disease, assuming a recessive mode of transmission; similarly, 48 SNPs were identified assuming a dominant mode and 53 SNPs in a continuous mode. When the 37 SNPs identified from the GLM recessive mode were used in a principal components analysis, the first principal component correctly predicted 46 of 50 cases and 47 of 50 controls. Among all the SNPs selected from GLMs for the three modes of transmission, 39 could be mapped to 1 of 33 genes. Many of these genes are involved in Q2 **various cancers, including *GASC1*, shown previously to be amplified in ESCCs, and *EPHB1* and *PIK3C3*. In conclusion, we have shown the feasibility of the Affymetrix 10K SNP array in genome-wide association studies of common cancers and identified new candidate loci to study in ESCC.** (Cancer Res 2005; 65(7): 1-5)

Introduction

Esophageal squamous cell carcinoma (ESCC) is one of the most common malignancies in the Chinese population. The standardized incidence rate of esophageal cancer in Shanxi Province, China is >100 per 100,000 person-years (1–3), although both incidence and mortality rates have declined slowly the past 10 years in this area (4). People in high-risk regions, such as Shanxi Province, are much more likely to develop this cancer than individuals residing in low-risk areas of the world. Within the high-risk regions, there is a strong tendency toward familial aggregation, suggesting that genetic susceptibility, in conjunction with environmental exposures, plays a role in the etiology of ESCC. ESCC is most likely a complex disease caused by mutations or risk alleles in multiple genes, each with a small contribution to overall risk. In the past several years, we and others have tried to identify susceptibility genes as well as biomarkers involved in ESCC, which can be used to screen high-risk populations in north central China, including genome-wide loss of heterozygosity testing, candidate tumor suppressor gene mutation testing, and analysis of expression arrays (5–10). The results from these studies indicate that several genes may play an important role in development of this tumor, but we have not yet found a classifier that can be used to screen high-risk populations.

Two approaches, linkage analysis and association studies, are commonly used to identify susceptibility genes involved in tumorigenesis. Linkage analysis involves genotyping of individuals from affected families, whereas association studies are done using subjects from population-based or family studies. In one example of such an association study, Sun et al. found that polymorphisms in apoptosis pathway genes *Fas* and *FasL* were associated with increased risk of developing ESCC (11). However, most studies were limited to the studies using a few single nucleotide polymorphism (SNP; refs. 12–14). It is estimated that SNPs occur one in every 1,000-bp nucleotides. Several genotyping studies on the chromosome-wide level using high-density SNPs have already been reported (15, 16). Recently, the GeneChip Mapping 10K Array for whole genome SNP analysis became available (Affymetrix, Inc., Santa Clara, CA) and a few initial reports of allelic imbalance or loss in cancer as well as cancer cell lines using the 10K SNP array have been published (17–23).

Here, we report the results of a pilot ESCC case-control study using the 10K SNP array. We had two primary and one secondary aims in this study. Our primary aims were to identify SNPs and genes that are associated with ESCC and to develop initial approaches appropriate for the analysis and interpretation of genome-wide association studies, including describing limitations and applications of such studies. Our secondary aim was to begin development of a classification method that combines multiple genotypes and environmental factors to predict susceptibility to ESCC.

Materials and Methods

Patients and controls

The study was approved by the institutional review boards of the Shanxi Cancer Hospital and the National Cancer Institute.

ESCC patients selected. Patients diagnosed with ESCC between 1998 and 2000 in the Shanxi Cancer Hospital in Taiyuan, Shanxi Province, People's Republic of China and considered candidates for curative surgical resection were identified and recruited to participate in this study. None of the patients had prior therapy and Shanxi was the ancestral home for all. After obtaining informed consent, patients were interviewed to obtain information on demographic and lifestyle cancer risk factors (smoking,

Requests for reprints: Maxwell P. Lee, Laboratory of Population Genetics, Center for Cancer Research, National Cancer Institute, Building 41, Room D702, 41 Library Drive, Bethesda, MD 20892. Phone: 301-435-8956; Fax: 301-435-8963; E-mail: leemax@mail.nih.gov.

alcohol drinking, and family history of cancer) and clinical data. We selected 50 males by identifying the first 25 with a positive family history of esophageal cancer and the first 25 without a family history of esophageal cancer by going down our roster ordered by study identification number.

Controls. Age-, sex-, and neighborhood-matched controls were selected and evaluated within 6 months of the case being diagnosed. The "neighborhood" in China refers to the residence blocks within communities. The ancestral home for all controls was also in Shanxi Province.

Biological specimen collection and processing.

Venous blood (10 mL) was taken from patients before surgery and from controls after interview. Germ line DNA was extracted and purified using standard methods.

GeneChip Mapping 10K Array.

The 10K SNP array provides comprehensive coverage of the genome for genotyping studies. Each array contained 11,555 biallelic polymorphic sequences randomly distributed throughout the genome, except for the Y chromosome. The median physical distance between SNPs is ~105 kb and the mean distance between SNPs is 210 kb. The average heterozygosity for these SNPs is 0.37, with an average minor allele frequency of 0.25. The algorithm used for making genotype calls was described previously by Affymetrix (24, 25).

Target preparation. DNA samples, including two control DNA samples from Affymetrix, were assayed according to the protocol (GeneChip Mapping Assay manual) supplied by Affymetrix. The procedure was similar to the one described previously (24). Briefly, a total of 250 ng germ line DNA were digested with *Xba*I and then ligated to *Xba*I adaptor before subsequent PCR amplification. All the steps mentioned above were carried on in the pre-PCR clean room. Cycling was conducted as follows: 95°C for 3 minutes followed by 35 cycles of 95°C for 20 seconds, 59°C for 15 seconds, and 72°C for 15 seconds. Final extension was done at 72°C for 7 minutes Q3 (DNA Engine Tetrad PTC-225, MJ Research). To evaluate PCR products, 3 μL of each PCR product were mixed with 3 μL of the 2 × gel loading dye on 2% Tris-borate EDTA gel and run at 120 V for 1 hour to check for the expected product (bands) between 250 and 1,000 bp. After purification and Q4 elution of the PCR products using Qiagen MinElute 96, quantification of purified PCR product was done using spectrophotometric analysis. A final 20 μg of PCR product were fragmented with DNase I. An aliquot of the fragmented PCR product was run on a 4% Tris-borate EDTA gel at 120 V for 30 minutes to 1 hour. Successful fragmentation was confirmed by the presence of a smear with the darkest region corresponding to 50 to 100 bp. The fragmented PCR product was end labeled with biotin and hybridized to the array. Arrays were incubated at 48°C for 18 hours in the Affymetrix GeneChip system hybridization oven. Microarrays were washed and stained in the GeneChip Fluidics Station 450 (Affymetrix) following the manufacturer's instructions.

Table 1. Summary of cases and controls and the 10K chip genotyping results

	ESCC (*n* = 50)	Controls (*n* = 50)
Risk factors		
Age (y), mean (SD)	58.2 (8.2)	57.8 (9.1)
Family history of esophageal cancer (yes)	0.50	0.14
Smoking (yes)	0.80	0.78
Alcohol drinking (yes)	0.38	0.14
Eating pickled vegetables (yes)	0.56	0.52
General results from the 10K SNP chip (%), mean (SD)		
Signal detection	98.9 (2.0)	99.0 (1.0)
SNP call	95.8 (0.9)	95.8 (2.2)
Genotype		
AA call	34.7 (0.5)	34.7 (0.5)
AB call	31.2 (0.7)	31.0 (0.7)
BB call	34.2 (0.5)	34.3 (0.4)

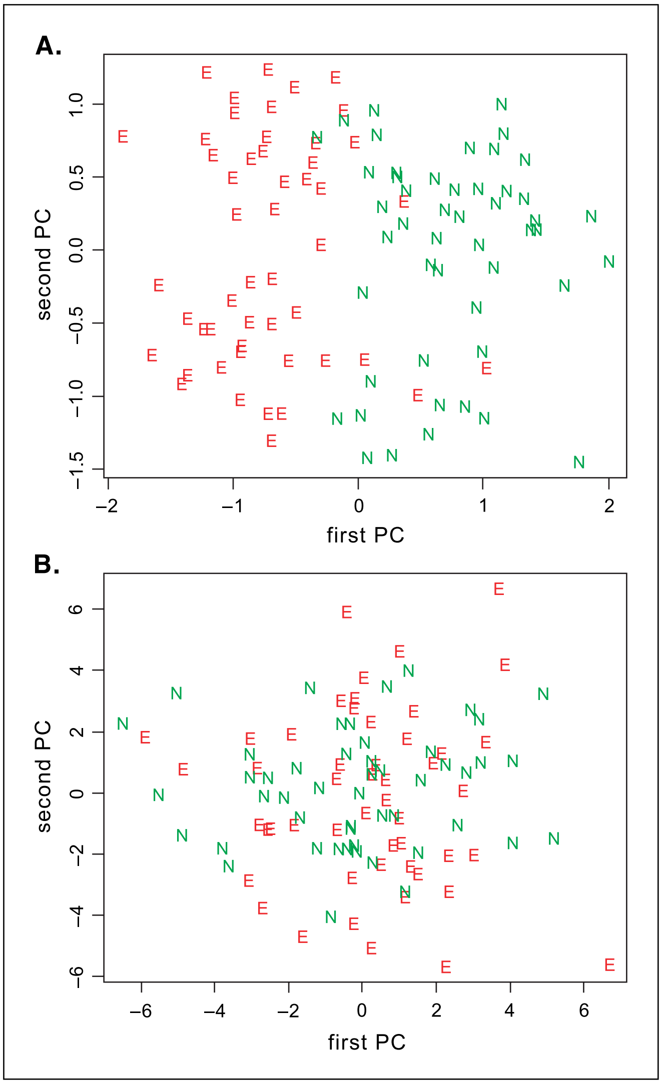

Figure 1. PCA analysis of cases and controls using multiloci genotypes. We used A allele as the recessive allele for genotype coding (AA, AB, BB) = (1, 0, 0). *E,* cases (*red*); *N,* controls (*green*). Details of the PCA can be found in Materials and Methods. *A, X axis* and *Y axis,* first and second principal components of PCA analysis using 37 SNPs from the 10K chip genotypes of 50 cases and 50 controls. *B, X axis* and *Y axis,* first and second principal components of PCA analysis using 3,369 SNPs (all SNPs with no missing genotype data) from the 10K chip genotypes of 50 cases and 50 controls.

Scanning and genotype generation. The 10K SNP arrays were scanned with the Affymetrix GeneChip Scanner 3000 using GeneChip Operating System 1.0 (Affymetrix). Data files were generated automatically. Genotype assignments (i.e., calls) were made automatically by GeneChip DNA Analysis Software 2.0 (Affymetrix). The genetic map used in the analysis was obtained from GeneChip Mapping 10K library files: Mapping10K_Xba131. "Signal Detection Rate" is the percentage of SNPs that pass the discrimination filter. "Call Rate" is the percentage of SNPs called on the array. The genotype calls are defined as AA, AB, or BB; "no call" means the SNP does not pass the discrimination filter.

Statistical analyses. All statistical analyses were developed using R and Splus packages. We applied the generalized linear model (GLM) implemented in the function GLM to evaluate the risk of each SNP that satisfied Hardy-Weinberg equilibrium at the significance level of $P > 0.01$. Three numerical coding schemes were used to represent genotypes: (*a*) (AA, AB, BB) = (1, 0, 0), (*b*) (AA, AB, BB) = (1, 1, 0), and (*c*) (AA, AB, BB) = (1, 0.5, 0). The first scheme corresponds to the assumption that allele A is recessive (equivalently, the allele B is dominant), the second scheme assumes that allele A is dominant (equivalently, the allele B is recessive), and the third scheme assumes a continuous mode.

GLM was applied to model the probability of being a case based on each SNP plus five potential explanatory variables, including $x1$ (family history positive, yes/no), $x2$ (alcohol use, yes/no), $x3$ (tobacco use, yes/no), $x4$ (pickled vegetable consumption, yes/no), and $x5$ (age, continuous):

$$\begin{aligned}\text{Prob} &= 1/(1+\exp(-f)), \text{where } f(x)\\ &= a + b^*\text{SNP} + b1^*x1 + b2^*x2 + b3^*x3 + b4^*x4 + b5^*x5.\end{aligned}$$

Three variables (age, smoking, and pickled vegetables) were insignificant for nearly all SNPs and were dropped from further consideration. Using a GLM for each SNP and the variables, we computed the P of the GLM based on the difference between null deviance D0 and residual deviance $D1$ using the χ^2 goodness-of-fit test. χ^2 statistic is $D0$-$D1$ with 3 *df.* To account for multiple comparisons, we used the Bonferroni-adjusted significance level to select our GLMs.

We used principal components analysis (PCA) to visualize similarity and variability among individuals. We applied PCA to each of the three numerical genotype coding schemes for all 100 case/control samples. The 100 samples were projected in the space defined by the first and second principal components. When case and control samples have two cluster structures in two principal components spaces, one or two principal components can be used to construct a classifier to separate cases and controls. The classifier was based on the genotyping of selected SNPs and its performance was evaluated for accuracy = (Tp + Tn) / 100, sensitivity = Tp / (Tp + Fn), and specificity = Tn / (Fp + Tn), where Tp and Tn are the numbers of true positives and true negatives and Fp and Fn are the numbers of false positives and false negatives. The odds ratio of the classifier is defined as Tp * Tn / [(50 − Tp) * (50 − Tn)]. Although developing and testing predictors using the identical same data is acknowledged to result in upward bias of predictor estimates (i.e., sensitivity, specificity, and accuracy), we calculated these values as a frame of reference only and not for clinical application without further confirmation (26).

Results and Discussion

In the present study, 50 male ESCC patients and 50 matched controls were examined using 10K SNP chips. Signal detection rates were high in both cases and controls (average of 98.9% and 99.1%, respectively), as were average SNP call rates (95.8% cases and 95.8% in controls; Table 1). The overall distributions of genotypes and allele frequency in the two groups are shown in Table 1.

T1

Based on National Center for Biotechnology Information (NCBI) Build 34, we summarized characteristics of the 11,555 SNPs and mapped these SNPs to chromosomes and genes. Thirty-four percent (3,947 of 11,555) of the SNPs were mapped in or near (within 1 kb of either 3′ or 5′ end) 2,187 different genes, including 108 SNPs in exons and 3,689 SNPs in introns. One hundred and thirty SNPs were removed because they could not be mapped to the human genome with NCBI Build 34. We removed another 953 SNPs that were homozygous in either case or control groups. We also removed 208 SNPs that did not satisfy Hardy-Weinberg equilibrium in the control group ($P < 0.01$). Following application of these filters, 10,264 SNPs remained for further analysis.

We first compared cases and controls for each of the 10,264 SNPs individually using multivariate analyses in the GLM assuming each of the three different modes of transmission described above (i.e., recessive, dominant, and continuous). Potential explanatory variables that might influence the analysis were adjusted for in the

Table 2. Summary of classification using multiple genotypes

Mode	No. SNPs	True negative	True positive	Accuracy	Sensitivity	Specificity	Odds ratio
Recessive	37	46	47	0.93	0.94	0.92	180.2
Recessive	31	47	43	0.90	0.86	0.94	96.2
Recessive	27	47	40	0.87	0.80	0.94	62.7
Recessive	22	44	42	0.86	0.84	0.88	38.5
Recessive	15	45	38	0.83	0.76	0.90	28.5
Recessive	10	42	38	0.80	0.76	0.84	16.6
Dominant	48	45	48	0.93	0.96	0.90	216
Dominant	41	45	45	0.90	0.90	0.90	81
Dominant	34	45	40	0.85	0.80	0.90	36
Dominant	28	45	41	0.86	0.82	0.90	41
Dominant	19	45	36	0.81	0.72	0.90	23.1
Dominant	10	34	39	0.73	0.78	0.68	7.5
Continuous	53	45	44	0.89	0.88	0.90	66
Continuous	42	44	43	0.87	0.86	0.88	45.1
Continuous	37	44	42	0.86	0.84	0.88	38.5
Continuous	26	46	37	0.83	0.74	0.92	32.7
Continuous	19	44	37	0.81	0.74	0.88	20.9
Continuous	9	42	32	0.74	0.64	0.84	9.3

NOTE: The 10 SNPs selected in the recessive coding schema: rs0560098, rs0952938, rs1014493, rs1104680, rs1559347, rs1824305, rs1986518, rs2168696, rs2419901, and rs4121091. The 10 SNPs selected in the dominant coding schema: rs0052911, rs0445077, rs0725517, rs0876658, rs1433114, rs1977656, rs2343264, rs2385456, rs2394122, and rs2401841. The nine SNPs selected in the dominant coding schema: rs0206847, rs0876658, rs0876660, rs0951538, rs1406121, rs1515366, rs1515367, rs1541290, and rs4121091.

GLM. Because 10,264 separate analyses were done, multiple comparisons were a major concern. We corrected for multiple comparisons using Bonferroni-adjusted significance levels, which, for 10,264 analyses, means that we accepted as significant only *P*s < 4.87187e−06 (which corresponds to a single test with α level of 0.05). Using multivariate GLMs with Bonferroni adjustment as described, we identified 37 statistically significant SNPs under the recessive transmission mode assumption, 48 SNPs for the dominant mode, and 53 SNPs assuming a continuous mode.

A secondary aim of this study is to develop in the future a method to predict individual risk of ESCC based on genotypes and explanatory variables. To begin approaching this aim, we combined the 37 SNPs selected from the recessive mode GLM to classify samples using PCA (Fig. 1*A*). With few exceptions, the cases and controls were clearly separated into two different clusters. As a comparison, we also did a PCA using all available SNPs in which there were no missing genotype data (*n* = 3,369 SNPs; Fig. 1*B*). It is clear that the PCA using all available SNPs resulted in no segregation between cases and controls, which serves to show that cases and controls came from the same population and that there were no major genotype differences between cases and controls at the population level. Given that there was good separation between cases and controls in the PCA using the 37 SNPs identified from GLM in the recessive mode, we developed a classifier to predict individual risk of esophageal cancer. Our classifier was defined by the first principal component (PC1), which contains weighed combinations of genotypes from these 37 SNPs. A person was classified as a case if PC1 was ≤0 or a control if PC1 was >0. Using PC1, we were able to correctly classify 46 of 50 cases and 47 of 50 controls. The accuracy, sensitivity, and specificity for this PCA classification were 0.93, 0.94, and 0.92, respectively (Table 2), and the odds ratio for being a case was 180.2. Similar results were also obtained when SNPs selected from the dominant or continuous mode GLMs were used (Table 2). We also did PCA loading analyses to assess discrimination when smaller numbers of the SNPs were used for classification. This analysis indicated that we could predict individual cancer risk using just 10 SNPs with an overall accuracy of 80%, sensitivity of 76%, and specificity of 84%; the odds ratio for

Table 3. Summary of SNPs and genes identified through whole genome association study using 10K chip

No.	Gene	SNP	Locuslink	Cytoband	Description
1	*ABCA4*	rs546550	24	1p22.1-p21	ATP-binding cassette, subfamily A member 4
2	*ACYP2*	rs2010461	98	2p16.2	Muscle-type acylphosphatase 2
3	*AKAP2*	rs723706	11217	9q31-q33	A-kinase anchor protein 2 isoform 2
4	*ALS2CR12*	rs1406121	130540	2q33.2	*ALS2CR12* gene product
5	*C13orf18*	rs1408195	80183	13q14.11	Chromosome 13 open reading frame 18
6	*DNAH8*	rs2050180	1769	6p21.31-p21.1	Dynein, axonemal, heavy polypeptide 8
7	*EFA6R*	rs1565138	23362	8p23.3	ADP-ribosylation factor guanine nucleotide factor 6
8	*EPHB1*	rs1515366, rs1515367	2047	3q21-q23	Ephrin receptor EphB1 precursor
9	*FLJ31810*	rs1021447, rs1343456	158038	9p21.1	
10	*FLJ34278*	rs1316607	283470	12q13.12	
11	*FLJ34922*	rs938298	91607	17q21.1	
12	*GALNT13*	rs707077	114805	2q24.1	UDP-*N*-acetyl-α-D-galactosamine:polypeptide *N*-acetylgalactosaminyltransferase 13
13	*GASC1*	rs1340513	23081	9p24	Gene amplified in squamous cell carcinoma 1
14	*GPR158*	rs1414045	57512	10p12.31	G protein–coupled receptor 158
15	*GRIN3A*	rs942142	116443	9q31.1	Glutamate receptor, ionotropic, *N*-methyl-D-aspartate 3A
16	*KIAA0802*	rs2143267	23255	18p11.22	
17	*KIAA1632*	rs1075906	57724	18q21.1	
18	*LOC374416*	rs536215			
19	*LOC374699*	rs727187			
20	*LOC376905*	rs1824305, rs1824306			
21	*LOC377255*	rs38055			
22	*MYH3*	rs876660 rs876658	4621	17p13.1	Myosin, heavy polypeptide 3, skeletal muscle, embryonic
23	*PACSIN1*	rs3845765, rs3800472	29993	6p21.3	Protein kinase C and casein kinase substrate in neurons 1
24	*PGRP-L*	rs959117	114770	19q13.2	Peptidoglycan recognition protein L precursor
25	*PIK3C3*	rs52911	5289	18q12.3	Phosphoinositide-3-kinase, class 3
26	*PKP4*	rs2108217	8502	2q23-q31	Plakophilin 4
27	*PLCB1*	rs2143267, rs2889786	23236	20p12	Phosphoinositide-specific phospholipase Cβ1 isoform b
28	*RAMP*	rs2014029	51514		RA-regulated nuclear matrix-associated protein
29	*SGCD*	rs256846	6444	5q33-q34	δ-Sarcoglycan isoform 2
30	*SLC9A9*	rs956062	285195	3q23	Solute carrier family 9
31	*SYNE1*	rs725467	23345	6q25	Nesprin 1 isoform α
32	*UBE2E2*	rs778480	7325	3p24.2	Ubiquitin-conjugating enzyme E2E 2
33	*XDH*	rs206847	7498	2p23-p22	Xanthine dehydrogenase

these 10 SNPs was 16.6 (Table 2). We also did permutation tests (1,000 tests) using randomly selected two thirds of the samples for training and one third of the samples for testing in PCA analysis. The permutation tests indicated that our PCA classification can be generalized. Hierarchical cluster analysis using the 37 SNPs selected from the GLMs in recessive mode was also able to classify cases and controls with similar performance (data not shown).

One alternative, and perhaps preferable, approach to reduce false-positive SNPs, beyond statistical adjustment, is to focus on SNPs that are in or near genes. When we combined results from our GLMs for all three modes of transmission, with information from NCBI Build 34 to identify SNPs in or near genes, we identified a total T3 of 39 SNPs in 33 genes (Table 3). Twenty of these 33 genes are named genes, many of which are involved in cancer. For example, *EPHB1* encodes a receptor tyrosine kinase and *PIK3C3* encodes a class 3 phosphoinositide-3-kinase. Receptor tyrosine kinase and phosphoinositide-3-kinase are common members of oncogenes. *GASC1* maps to 9p24, a region frequently amplified in ESCCs (27). Yang et al. (27) cloned *GASC1*, which stands for gene amplified in squamous cell carcinoma-1, and showed that *GASC1* was overexpressed in several cell lines. GASC1 protein contains 2 PHD finger motifs and a PX domain. PHD finger motifs are zinc finger-like sequences found in nuclear proteins that function in chromatin-mediated transcriptional regulation and are present in some oncogenes. The SNP rs951998 was also identified by GLM, although it was not included in Table 3 because it is located 25 kb upstream of *CDK8* and 241 kb downstream of *RNF6*. Interestingly, somatic mutations in *RNF6* in ESCC tumor samples were reported previously (8).

Herein, we have described our initial efforts using genome-wide SNP arrays applied to germ line DNA in a population-based epidemiologic case-control association study to explore genetic susceptibility to ESCC. We have addressed two of the major methodologic concerns in such studies, potential confounding and multiple comparisons, by adjusting for numerous potential confounders in our statistical models and by accepting SNP associations as real only under very stringent and conservative statistical conditions. The SNPs we identified in our GLMs as associated with ESCC seem to be robust in their ability to separate cases from controls. Each of the various discriminatory methods we applied—GLM with different modes of transmission, PCA with various number of SNPs, and hierarchical clustering—all distinguished cases from controls. We are encouraged that PCA using multiloci genotyping may provide a valuable new tool for assessing risk of developing ESCC at the level of the individual. In the meantime, several different but complementary approaches remain to be pursued: further family-based linkage analysis will be used to confirm a subset of the loci that are genetically linked to ESCC; additional genotyping using higher-density arrays, such as Affymetrix 100K chip, and more detailed examination of SNPs across the 33 loci identified here will permit identification of haplotype block structures that will further refine the mapping and cloning of genes important for the etiology of ESCC; case-control studies involving more subjects will permit testing and refinement of SNP profiles for risk prediction; and molecular genetic studies of the 33 genes reported here will provide additional evidence for the role of these genes in ESCC.

Acknowledgments

Received 9/7/2004; revised 11/17/2004; accepted 1/18/2005.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank Jenny Kelley for critical reading of the article.

References

1. IARC. Alcohol drinking. IARC monographs on the evaluation of carcinogenic risks to humans, 44. Lyon: IARC; 1988. p. 153–246.

2. Li JY. Epidemiology of esophageal cancer in China. Natl Cancer Inst Monogr 1982;62:113–20.

3. National Cancer Control Office. Investigation of cancer mortality in China. Beijing: People's Health Publishing House; 1980.

4. Qiao YL, Hou J, Yang L, et al. The trends and preventive strategies of esophageal cancer in high-risk areas of Taihang Mountains, China. Zhongguo Yi Xue Ke Xue Yuan Xue Bao 2001;23:10–4.

5. Hu N, Roth M, Polymeropolous M, et al. Identification of novel regions of allelic loss from a genome-wide scan of esophageal squamous cell carcinoma in a unique high-risk population. Genes Chromosomes Cancer 2000; 27:217–28.

6. Roth MJ, Hu N, Emmert-Buck MR, et al. Genetic progression and heterogeneity associated with the development of esophageal squamous cell carcinoma. Cancer Res 2001;61:4098–104.

7. Hu N, Huang J, Goldstein AM, et al. Frequent inactivation of the TP53 gene in esophageal squamous cell carcinoma from a high-risk population, China. Clin Cancer Res 2001;7:883–91.

8. Lo HS, Hu N, Gere S, et al. Identification of somatic mutations of the RNF6 gene in human esophageal squamous cell carcinoma. Cancer Res 2002;62:4191–3.

9. Su H, Hu N, Shih J, et al. Gene expression in esophageal squamous cell carcinoma reveals highly consistent and cancer family history-related molecular profiles. Cancer Res 2003;63:3872–6.

10. Hu N, Li WJ, Su H, et al. Common genetic variants of TP53 and BRCA2 in esophageal cancer patients and healthy individuals from low and high risk areas of northern China. Cancer Detect Prev 2003;27:132–8.

11. Sun T, Miao X, Zhang X, Tan W, Xiong P, Lin D. Polymorphisms of death pathway genes FAS and FASL in esophageal squamous-cell carcinoma. J Natl Cancer Inst 2004;96:1030–6.

12. Zhang J, Li Y, Wang R, et al. Association of cyclin D1 (G870A) polymorphism with susceptibility of esophageal and gastric cardiac carcinoma in a northern Chinese population. Int J Cancer 2003;105:281–4.

13. Hao B, Wang H, Zhou K, et al. Identification of genetic variants in base excision repair pathway and their associations with risk of esophageal squamous cell carcinoma. Cancer Res 2004;64:4378–84.

14. Wu MT, Lee JM, Wu DC, et al. Genetic polymorphisms of cytochrome p4501A1 and esophageal squamous-cell carcinoma in Taiwan. Br J Cancer 2002; 87:529–32.

15. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander E. High-resolution haplotype structure in the human genome. Nat Genet 2001;29:229–32.

16. Ozaki K, Ohnishi Y, Iida A, et al. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat Genet 2002; 32:650–4.

17. Phillips MS, Lawrence R, Sachidanandam R, et al. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet 2003; 33:382–7.

18. Zhao X, Li C, Paez JG, et al. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 2004;64:3060–71.

19. Janne PA, Li C, Zhao X, et al. High-resolution single-nucleotide polymorphism array and clustering analysis of loss of heterozygosity in human lung cancer cell lines. Oncogene 2004;23:2716–26.

20. Wong KK, Tsang YTM, Shen J, et al. Allelic imbalance analysis by high-density single-nucleotide polymorphic allele (SNP) array with whole genome amplified DNA. Nucleic Acids Res 2004;32:e69.

21. Sellick GS, Garrett C, Houlston R. A novel gene for neonatal diabetes maps to chromosome 10p12.1-p13. Diabetes 2003;52:2636–8.

22. Middleton FA, Pato MT, Gentile KL, et al. Genome-wide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. Am J Hum Genet 2004;74:886–97.

23. Zhou X, Li C, Mok SC, Chen Z, Wong DTW. Whole genome loss of heterozygosity profiling on oral squamous cell carcinoma by high-density single nucleotide polymorphic allele (SNP) array. Cancer Genet Cytogenet 2004;151:82–4.

24. Kennedy GC, Matsuzaki H, Dong S, et al. Large-scale genotyping of complex DNA. Nat Biotechnol 2003;21: 1233–7.

25. Liu WM, Di X, Yang G, et al. Algorithms for large-scale genotyping microarrays. Bioinformatics 2003;19: 2397–403.

26. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95: 14–8.

27. Yang ZQ, Imoto I, Fukuda Y, et al. Identification of a novel gene, GASC1, within an amplicon at 9p23-24 frequently detected in esophageal cancer cell lines. Cancer Res 2000;60:4735–9.